Simone Licciardi
Politecnico di Milano, Milan, IT
simone.licciardi@outlook.com
+39 342 020 9779

Adaptive Learning Rates via Continuous-Time min max Optimization Model

Introduction

Recently, min max optimization problems [1] have attracted significant interest from fields such as ML [2, 3], Control Theory [4], Network Design [5, 6] and Game Theory as a universal model for adversarial zero-sum games. To solve them, the most common algorithms are Gradient Descent Ascent (`GDA`) and its variants [1, 7]. These algorithms can be unstable due to the numerical problem's *stiffness*, termed *time-scale separation* [8] in optimization: when one directional update dominates the other. Some heuristic solutions have been designed, with the most advanced being `NeAda` [8] and `TiAda` [9].

Simultaneously, recent works on continuous-time modeling of Gradient Descent (`GD`) have shifted away from theory over the Negative Gradient Flow and towards better models for the discrete-step dynamics of `GD` [10, 11, 12]. This effort explained empirical observations (e.g. Implicit Regularization [11], Edge-of-Stability behavior [13]), produced practical results (adaptive learning rate schedules based on Quasi-Newton approximations of the Hessian), and models such as the Principal Flow (`PF`) [14]. While works on continuous flow for min max optimization delivered promising results through regularization [15], the development of techniques to mitigate the time-scale separation remains little investigated.

Objectives

The primary goal of this proposal is to study the separation of time-scales by using continuous-time flows, and determine theory-driven adaptive learning rates for `GDA` as a consequence. To achieve this, we want to 1. adapt continuous-time models (such as `PF`) to min max problems, 2. utilize these models to investigate the cross-influence between the primal and dual variables, 3. build on this to make the time-scale constant [8] explicit as a dynamical quantity of the model. As a downstream task, 4. create a computationally tractable schedule for adaptive learning rates that.

The main challenge will be to account simultaneously for the mitigation of the Edge-of-Stability behavior (Requirement **A**), which motivates the `PF` model, and the time-scale concerns (Requirement **B**), which motivates this analysis. Specifically, **B** enforces a constraint on the ratio of learning rates for the primal and dual variables, preventing us from choosing learning rates independently for each direction.

Methodology

The first approach is to build on [12, 15]. We use Backward Error Analysis [16] to find a continuous flow model, then find a surrogate for the time-scale constant L^2 from [8] through the hessian of the objective function, and finally make it explicit through the dynamical model. To construct the learning rate schedules, we start with the optimal ones deduced from `PF` in [14] to ensure (**A**) and then rescale them to fit the time-scale from the dynamical, as to guarantee (**B**). This can be made tractable with Quasi-Newton approximations of the Hessian, like in [14].

The alternative approach, suggested by the management of *stiffness* in numerical analysis, is to consider a different `GDA` model: the vanilla (min max) Implicit Gradient Flow (`IGF`) [11], with added stochastic noise. We would then model this in continuous time through Stochastic DEs. This is a viable approach: noise allows us to adapt `GDA` timesteps until they fit the time-scale constant, satisfying (**B**), and yet escape local minima through stochastic noise, resolving `IGF` being attracted to them [12].

Relevance

Stable min max algorithms are essential. Thus, here are some of the most relevant applications of improved optimization schedules: 1. to stabilize GANs training and avoid mode collapse, 2. to produce effective adversarial training defense [17, 18], where SOTA (termed `FAT`) relies on `GDA`. Better adversarial defence ensures deep learning reliability for critical applications such as energy and infrastructure. 3. to implement robust decision-making in an embedded and dynamical setting.

References

[1] Meisam Razaviyayn et al. "Nonconvex Min-Max Optimization: Applications, Challenges, and Recent Theoretical Advances". In: *IEEE Signal Processing Magazine* 37.5 (Sept. 2020), pp. 55–66. ISSN: 1558-0792. DOI: 10.1109/msp.2020.3003851. URL: http://dx.doi.org/10.1109/MSP.2020.3003851.

[2] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML]. URL: https://arxiv.org/abs/1406.2661.

[3] David Madras et al. *Learning Adversarially Fair and Transferable Representations*. 2018. arXiv: 1802.06309 [cs.LG]. URL: https://arxiv.org/abs/1802.06309.

[4] D.M. Raimondo et al. "Min-max Model Predictive Control of Nonlinear Systems: A Unifying Overview on Stability". In: *European Journal of Control* 15 (Dec. 2009). DOI: 10.3166/ejc.15.5-21.

[5] Ramy Gohary et al. "A Generalized Iterative Water-Filling Algorithm for Distributed Power Control in the Presence of a Jammer". In: *Signal Processing, IEEE Transactions on* 57 (Aug. 2009), pp. 2660–2674. DOI: 10.1109/TSP.2009.2014275.

[6] Meisam Razaviyayn, Mingyi Hong, and Zhi-Quan Luo. *Linear Transceiver Design for a MIMO Interfering Broadcast Channel Achieving Max-Min Fairness*. 2012. arXiv: 1208.6357 [cs.IT]. URL: https://arxiv.org/abs/1208.6357.

[7] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. *A Unified Analysis of Extra-gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach*. 2019. arXiv: 1901.08511 [math.OC]. URL: https://arxiv.org/abs/1901.08511.

[8] Junchi Yang, Xiang Li, and Niao He. *Nest Your Adaptive Algorithm for Parameter-Agnostic Nonconvex Minimax Optimization*. 2022. arXiv: 2206.00743 [math.OC]. URL: https://arxiv.org/abs/2206.00743.

[9] Xiang Li, Junchi Yang, and Niao He. *TiAda: A Time-scale Adaptive Algorithm for Nonconvex Minimax Optimization*. 2022. arXiv: 2210.17478 [math.OC]. URL: https://arxiv.org/abs/2210.17478.

[10] Omer Elkabetz and Nadav Cohen. *Continuous vs. Discrete Optimization of Deep Neural Networks*. 2021. arXiv: 2107.06608 [cs.LG]. URL: https://arxiv.org/abs/2107.06608.

[11] David G. T. Barrett and Benoit Dherin. *Implicit Gradient Regularization*. 2022. arXiv: 2009.11162 [cs.LG]. URL: https://arxiv.org/abs/2009.11162.

[12] Mihaela Claudia Rosca. *On discretisation drift and smoothness regularisation in neural network training*. 2023. arXiv: 2310.14036 [stat.ML]. URL: https://arxiv.org/abs/2310.14036.

[13] Jeremy M. Cohen et al. *Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability*. 2022. arXiv: 2103.00065 [cs.LG]. URL: https://arxiv.org/abs/2103.00065.

[14] Mihaela Rosca et al. *On a continuous time model of gradient descent dynamics and instability in deep learning*. 2023. arXiv: 2302.01952 [stat.ML]. URL: https://arxiv.org/abs/2302.01952.

[15] Mihaela Rosca et al. *Discretization Drift in Two-Player Games*. 2021. arXiv: 2105.13922 [stat.ML]. URL: https://arxiv.org/abs/2105.13922.

[16] Ernst Hairer et al. "Report 14/2006: Geometric Numerical Integration (March 19th – March 25th, 2006)". In: *Oberwolfach Reports* 3 (Dec. 2006). DOI: 10.4171/OWR/2006/14.

[17] Weimin Zhao, Sanaa Alwidian, and Qusay H. Mahmoud. *Adversarial Training Methods for Deep Learning: A Systematic Review*. 2022.

[18] Aleksander Madry et al. *Towards Deep Learning Models Resistant to Adversarial Attacks*. 2019. arXiv: 1706.06083 [stat.ML]. URL: https://arxiv.org/abs/1706.06083.